NCI Center for Bioinformatics
Informatics Seminar Series

—

Three Aspects of Computational Bioinformatics at the San Diego Supercomputing Center

9:00 until Noon
September 17, 2001

—

Neuroscience Bldg, 6001 Executive Blvd.
Conf. Room A1/A2

—

"Challenges and Opportunities of Structural Genomics"

Philip E. Bourne, PhD
National Biomedical Computational Resource
San Diego Supercomputer Center

Ilya Shindyalov, B.V.B. Reddy of SDSC and Philip E. Bourne of SDSC and the Department of Pharmacology at UC San Diego work on ways to characterize structure information for use in protein engineering and drug target selection. They have developed a structure comparison technique called Combinatorial Extension (CE) that compares the 3-D structures of proteins and operates on computers part of the national grid as well as in individual laboratories. CE has importance for finding both protein function and evolutionary relationships among proteins. Results from this work in addressing the Paracelsus challenge for protein engineering and new discoveries regarding protein function will be presented. Further detail son CE can be found at http://www.npaci.edu/envision/v16.3/ce.html.

Shindyalov, Reddy and Bourne have also developed a template cluster based modeling approach to model short peptide structures useful for drug designing studies. The method has been tested successfully for designing analogues of organic compounds with the ability to mimic peptides with arterial growth-stimulating activity. These and other tools will be highlighted in this presentation.

—

"Macromolecular Pattern Recognition and On-line Access to Molecular Biology Tools"

Michael Gribskov
National Biomedical Computational Resource
San Diego Supercomputer Center

A key challenge for modern biology is to keep track of large, complex sets of data, such as the human genome and to relate these large data sets both to other public data and to locally developed information resources. Gribskov's group at NBCR and the Biology Department at UCSD are developing applications for recognizing motifs and for federated database query tools, all accessible through the Web. This research infrastructure has been highlighted in several large forums, and is being submitted for production by the larger community [http://www.npaci.edu/envision/v16.3/nbcr.html]. These efforts include the Homophila database. The purpose of the Homophila database is to utilize the sequence information of human disease genes from the Online Mendelian Inheritance in Man (OMIM) database in order to determine if

sequence homologs of these genes exist in the current *Drosophila* sequence database FlyBase . We find that 77% of human disease gene associated sequences in OMIM have strong matches (e < 10 -10) to one or more sequences in the *Drosophila* database.[http://homophila.sdsc.edu/]. A new project in its preliminary stages focuses on the use of RNA splicing isoforms as a tool to aid in the diagnosis and treatment of human cancer. This project, a collaboration with Xiang-Dong Fu of UCSD and Michael Zhang of Cold Spring Harbor Laboratory, will create a database of RNA splicing isoforms, develop microarray methods for measuring the abundance of specific spliced forms, and ultimately develop predictive software for use in diagnosis and treatment. This project is now moving from a preliminary, proof-of-concept stage into actual construction of database of RNA splicing. These and other projects will be highlighted in this presentation.

—

"Model-Based Integration of Scientific Data: Database Mediators Meet Knowledge Representation"

Bertram Ludaescher
San Diego Supercomputer Center

Database mediator systems integrate heterogeneous data sources by defining a common virtual database view in a suitable high-level query language. In the MIX (Mediation of Information using XML) Project, a collaboration of researchers from SDSC/DICE and the UCSD database lab, technologies for specifying, rewriting, and evaluating XML-based virtual views have been developed. These technologies are being applied to scientific applications in a variety of domains ranging from Earth Systems Sciences to Neurosciences. Experience with highly complex, so-called "multiple-world" integration scenarios from these domains has shown the limitations of current syntax and structure-oriented "pure XML" approaches. Explicit modeling of domain semantics seems necessary for bridging the information gaps between the apparently disjoint multiple worlds. Domain Maps are ontologies defined by domain scientists and "mediation engineers" that provide the semantic glue for mediation and query processing across multiple source worlds. By formalizing domain maps and source models in description logics and expressing integrated views as logic rules, automated deduction techniques can be employed to assist the domain scientists and mediation engineers during the source model design (intra-model consistency), integrated view definition (inter-model
consistency), and deployment (e.g., for semantic query optimization).

Our research is driven by an ongoing collaboration between researchers from SDSC and NCMIR, which provides ideal, real-world test cases for complex, multiple world mediation scenarios from the Neuroscience domain. We illustrate our approach using a prototype for Knowledge-based Integration of Neuroscience Data (KIND). (joint work with A. Gupta and M. E. Martone)